

Implementation and Validation of Language Proficiency and Interpreter Readiness Tests

Jean Turner, California Endowment / Monterey Institute

I am at the Monterey Institute, and the work that I have done on the test that I'm going to be talking about has been as a consultant to projects that were funded by the California Endowment. Those projects were funded through several different entities, so we weren't really sure how to characterize where these tests came from; therefore we decided to give the credit to the funding agency, which was the California Endowment. I am suggesting that we refer to them as the California Endowment Tests, because there are a lot of AKAs or aliases.

I'm going to describe these tests briefly, because there is actually a whole suite of tests. I'd like to note that some people refer to suites of tests as test batteries, and I think that that is an appropriate word for a suite of tests. I think it's something we should consider as we're thinking about a certification effort. Is it going to be a test battery, or is it going to be a single test?

So this test program that I'm going to be talking about includes four tests in several different language combinations. There are language proficiency tests in Spanish, Cantonese and Hmong, and there's also an English-language proficiency test. There's something that's called an Initial Interpreter Readiness Test and something that's called a Final Interpreter Readiness Test. I'm sharing this terminology with you because this is the terminology being used in various reports that are now being written about the validation work on the tests. I think it is helpful to begin to think of the tests by these names.

First, I'll describe briefly the language proficiency tests. The Spanish, Cantonese and Hmong language proficiency tests were designed and developed by Claudia Angelelli, which is why some people refer to them as Claudia's tests. They were developed with funding by the California Endowment for the Connecting Worlds Partnership or Consortium. There was also an English language proficiency test, which was developed by my colleague Renee Jourdenais and me. It's based on the design specifications for the other language proficiency tests, but it's a little bit different, because of the requirements for English being a little bit different than the requirements for some of the other languages, specifically Hmong. There is not a reading or writing component to the Hmong language proficiency test. The original purpose of these tests, since they were developed for the Connecting Worlds Consortium, was to determine whether the individuals who presented themselves for this training had the language proficiency in both their target language and English to engage in training.

The specific features of these tests include that each one of them has six tasks; those tasks are designed to measure listening and in some cases reading and speaking. The test content is delivered through audiotape or CD. There's a very small writing component on the Spanish, Cantonese and English language proficiency tests. There is no writing and there is no reading on the Hmong language proficiency test. I've listed a couple of example tasks, not all of them, because there's not time for that. I don't know that we really need to go into that depth anyway, but take as an example the reading task. Examinees might read a label on a medicine bottle and then answer

comprehension questions. By the way, all of the responses are oral and are recorded for later scoring. An example of a listening task might entail listening to a patient's description of a medical condition and answering comprehension questions regarding that. Another possible example is a speaking task where one listens to a health care provider describe a medical procedure and then, in the same language, simply explain that to a patient who may not have understood the register. It's a thirty-five minute administration time. There's scoring by two trained raters.

The Interpreter Readiness tests were developed in three language combinations: Spanish-English, Cantonese-English, and Hmong-English. Again, they were designed by Claudia, so these too are sometimes referred to as Claudia's tests or the Connecting Worlds tests, or the *Hablamos Juntos* tests, which I will get to in a few moments, or LESA, or the California Endowment tests. As I indicated previously, I'll refer to the suite of tests as the California Endowment tests. The original purposes of the Initial Interpreter's Readiness was as a screening or a pre-test prior to interpreter training, with scores to be used to determine whether someone had the requisite interpreting skills for engaging in the medical interpreting training that was offered. The final Interpreter Readiness test was designed to be given following training, to measure whether people had the requisite language and interpreting skills to function specifically as a community medical interpreter.

Here are some of the features of these tests. I'm going to call them the IR tests because I only have twenty minutes, and Initial Interpreter Readiness and Final Interpreter Readiness is going to take me a long time to say every time I have to say it. So both IR tests have four short video-taped segments of simulated interactions between a health care professional and a patient. The patient is Spanish-speaking, or Cantonese-speaking or Hmong-speaking, and the health care professional is English-speaking. The examinee takes the role of the interpreter viewing the role play or the scenarios. There are pauses for that person to do interpreting. The responses are recorded for scoring, which is done by two trained raters. Scoring is done by people who are proficient in both English and Spanish, and there is a scoring rubric that presents the script for the scenarios, organized by turns. And then there are critical components of the turns that are scored separately, distinctly for accuracy: one point for accurate, no point for inaccurate, as judged by the trained raters. Not all turns are scored, and not all of the information in each turn is scored.

Here are some of the initial IR tasks. By the way, there are four scenarios, but it's the same patient engaging in several different tasks, things that patients do. So first, a patient makes and verifies an appointment, changes an appointment, has the appointment with a doctor, and then makes an urgent care appointment request and interacts with several health care professionals in that. That one is pediatric, by the way. In the final IR tasks, the patient makes an appointment and sees a specialist. There is sight translation of a consent form, and then an admission process in the third scenario, and then the fourth is a follow-up appointment. Perhaps interestingly to some of you, the sight translation in the trials that were done of the Hmong-English version of the test was not performed by any of the candidates.

Those are the tests. Those were developed with funding by the California Endowment and by Claudia. I was a consultant on that project. My involvement in the first and following phase of the implementation and validation work was as project manager for that effort, separate and distinct from the test development process. The

goals of this first phase of the validation work that was completed in 2006 were to support the implementation of the test. The tests had been developed, but they were not in use.

We presented the tests to the five organizations that had contracted them, and trained people at the sites in how to administer the tests. We identified and trained a pool of raters at each one of the sites, or a rater at each one of the sites. We also determined rater reliability, and we had planned to also investigate the impact of training on test performance and the impact of the amount of experience on test performance. This is the information about the reliability in general for the tests that were given. We had thought that we would be able to do this battery of testing with twenty to twenty-five people at each of the sites. In fact, we ended up with a very different configuration of people. For example, there was a total of fifty-one people who did the English test, the English language proficiency test, and only six people who did the English-Cantonese Final Interpreter Readiness Test. The rater reliability ranged from 0.55 for the Cantonese Final IR, to 0.84, for Spanish Final Interpreter Readiness test, I believe.

For the impact of training, we identified people within our rather small number of participants who had forty hours of training or less, and fifty hours or more. We wanted to know whether there was a difference in their performance on the test, because we thought the results would give us an idea of whether the tests were measuring the appropriate skills, knowledge and abilities. We found that on both the Initial and the Final Interpreter Readiness Tests there were no significant differences in performance. We also investigated whether amount of experience would have an impact on test performance. We thought people who had more experience would do better on the tests if the tests really measure valued and valuable skills. We identified people who had had one year or less of experience in our sample and people who had one year or more. We found on the Initial Interpreter Readiness Test that there was a statistically significant difference. Some of you might say, "Yeah!" Whoops! It was the less experienced people who did better! I have to say, I don't think this finding is a reflection on the test. My feeling is that the test is soundly designed, and we simply had huge problems in getting people to come and take the tests. I am quite certain that those people who did participate are not representative of the typical audience for the test. There was no significant difference for experience on the Final Interpreter Readiness Test.

Because of the very small sample for the different combinations of tests, and because there are questions that remain, we still need to collect validity evidence for the tests. There is a second phase of validation work that has been funded. I'm not going to go into the details of that; instead I would like to share some more thoughts on the validation work that has been done. First, the empirical basis for a test may provide convincing evidence of its content validity, but collecting evidence to support the usefulness of a test, that is, its value for screening purposes, is challenging, as is investigating the meaningfulness of test scores. What is a good score and at what score can one be assumed to be qualified or a master? We tried to investigate these questions through the differential group studies and found that for these people who were willing to come in and take the tests, there was no apparent difference. So when thinking about recruiting field test participants, we need to consider how a representative sample can be obtained. I ask myself, "Who likes to take tests? Who is willing to come in and do two hours of testing for sport? And are those people typical?" I'd like to suggest that they're not, which is a huge problem when we're investigating the validity of a test. We have to be very careful to design the research so that we do have access to people

who are typical of the intended audience. I don't think the people that we recruited were, even though they were offered fifty dollars and our admiration and esteem. Is fifty dollars enough to alleviate test trepidation for most people? I don't think so. There are many other theoretical and logistical and practical concerns related to designing and conducting validity research that one has to think about, ponder on, work collaboratively to solve before beginning the really important part of the development of the test.

Q: You mentioned the Cantonese test. There is a written part, right? Is the language itself Cantonese or Mandarin?

A: You know I believe it's Cantonese. At this level of questioning, we would have to look at the test itself. It was reviewed by a person as part of the validity work that I was involved in. We reviewed all of the tests and I was told that it was Cantonese. And I have been told by the people who developed it that it is Cantonese. The fact that it was reviewed by someone who wasn't involved in the development and who also assured me that it was Cantonese was comforting, because when we had the Hmong test similarly reviewed by someone who had not been one of the informants in its creation, we discovered that in fact it was not the variety of Hmong that we had been informed that it was, not precisely the color of Hmong that we had been informed.

Q: You mentioned that the folks who had less experience did better on this test. How was experience defined?

A: Experience was defined as they answered on a questionnaire saying, "How long have you been working as an interpreter?" And if they identified more than a year, we put them in one group.

Q: Is this regardless of actual experience within that year, because I've been doing it for a year, but I've only interpreted five times.

A: Just how long they've been doing it. The question was, "How long have you been doing interpreting?" People said, "I haven't done it." People said, "Two months." People said, "I've been an interpreter since I was twelve." So just simply on the basis of this self-report data, we tried to sort people out into people who had had more than a year and people who had less. And that could be one of the reasons that we came up with funky, unexpected outcomes.

Q: Along that line, did you ask any questions in regard to the fact that they had formal training or not? I'm just wondering whether that may have made a difference.

A: Yes. We did an analysis for training and people who had had forty hours of training or more were put in one group, and people who had forty hours or less were in another. Forty hours was defined because that is typically the length of the training program that is offered by the Connecting Worlds. By the way, during the eighteen months that we were working on this project, that training wasn't offered. So we couldn't actually do this project pre-post as we had intended to do. So we did have a group of forty hours or less and more than forty hours. More than forty hours might have been "I have a bachelor's degree. I did a certificate. I did this training, plus I've also had another six months of

study somewhere.” And I would say that the people who identified themselves as having forty hours or less had done the training that was offered by whichever organization they were being tested at.

Q: Did they fare better or worse?

A: They were the same. No difference in the groups.

Q: I have one comment and then a question. In the early stages of the federal exam, back in 1980, we did a study of the first group that took the test and found a negative correlation between years of experience and passing the test, and no, there’s nothing wrong with the test. What’s wrong is that people were working who did not have linguistic and interpreting capability at the level required and probably didn’t understand the goals of interpreting, fidelity, concept verbatim, etc, etc. They didn’t understand their ethical roles, and that’s certainly a validation of what you found.

A: It’s possible. I have a feeling. I hate to even call them findings, because they’re not really findings. They are based on a very, very small sample of people who were recruited. Thank you very much for recruiting them. You know, people worked very hard at these sites to get people in so that we could try these tests out, but those people are not typical of the general group of community medical interpreters that would come into those offices.

Q: Well, these were real findings, and we had like two-hundred-fifty people. And it was really shocking to everyone at the time, and we were all aghast, and we had to figure out what was going on.

A: I appreciate that, and as we do more research on this test, if we continue to find this profile, I think it will have really important implications for attention to assistance for community medical interpreters.

Q: Right. Exactly. And then, number two, my question is you found no increase, or significant increase in the testing after taking the Connecting Worlds?

A: None of the participants actually took the training. That was our original design, to do this as a pre-post, to have people take the appropriate language proficiency test and the English proficiency test, and the Initial Interpreters Readiness Test, and then engage in the training, and then do the post test. But that did not happen at any of the sites, because during the eighteen months we were engaged in the project, none of the sites offered that training because they didn’t have the funding at that point in time to do it.

Q: But you reported it as a finding? Did I read it wrong? There was no significant increase.

A: That is for experience and for amount of training, but it wasn’t the training that was provided by these agencies. It was the self-report training. Thank you for clarifying that.

Q: If that were true, then there would have to be a real study of the tasks, skills and abilities being taught in that program and if they matched the testing, which I wonder about.

A: In fact, when Claudia was engaged to develop the test, it was intended that it would be used with the Connecting Worlds training. But the Connecting Worlds training was not done when this test was done. So in fact, she had to kind of imagine and hypothesize, and I understand that it is now a curriculum which is used in different ways at the different organizations depending on the type of people that come into these five different organizations. I'm sorry about that lack of clarity. There was no training done with these participants as part of this study.

Q: On the Hmong interpreters, was your experience that you were just not able to get anybody to come and take the test? Was that what happened?

A: For the validation project, we didn't work with the Hmong test. We focused on the Spanish and Cantonese tests because in the initial development of the Hmong test, I think that there were some issues. I wasn't a complete insider on that first project, but I did work with the Hmong group in the orientation and training we did to introduce the tests to those people. And I believe that the nine interpreter trainers who were working for one of the organizations and had contributed to the development of that test also took the test. I think that there were a couple of other people who took the test as part of this preliminary pilot, and none of them completed the screening. And when I asked them about it, because I was the person who was working with that group, they just said, "Oh, we don't do that."

Q: My comment would be that in Madison, when we actually started to have an assessment for our Hmong interpreters, the same way we were doing for Spanish interpreters, all of them refused to come and take the test. As funny as it sounds, it really meant something. So I would like us as we think about this issue of certification to keep in mind that we're not going to be able to go and drag people from their homes to come and take the certification. So when I was talking about the end goal being language access for limited English-speaking people, we really need to keep those issues in mind. If all my Hmong interpreters in Madison refused to take the test, then I can have as many tests as I want, it's not going to happen. And they told me they just don't think the test is a good way to assess their skills and that they were insulted by even the idea, thus they were not going to come. Just food for thought.